mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding

Year: 2023 | *Citations: 154* | *Authors: Jiabo Ye, Anwen Hu, Haiyang Xu*

Abstract

Document understanding refers to automatically extract, analyze and comprehend information from various types of digital documents, such as a web page. Existing Multi-model Large Language Models (MLLMs), including mPLUG-Owl, have demonstrated promising zero-shot capabilities in shallow OCR-free text recognition, indicating their potential for OCR-free document understanding. Nevertheless, without in-domain training, these models tend to ignore fine-grained OCR features, such as sophisticated tables or large blocks of text, which are essential for OCR-free document understanding. In this paper, we propose mPLUG-DocOwl based on mPLUG-Owl for OCR-free document understanding. Specifically, we first construct a instruction tuning dataset featuring a wide range of visual-text understanding tasks. Then, we strengthen the OCR-free document understanding ability by jointly train the model on language-only, general vision-and-language, and document instruction tuning dataset with our unified instruction tuning strategy. We also build an OCR-free document instruction understanding evaluation set LLMDoc to better compare models' capabilities on instruct compliance and document understanding. Experimental results show that our model outperforms existing multi-modal models, demonstrating its strong ability of document understanding. Besides, without specific fine-tuning, mPLUG-DocOwl generalizes well on various downstream tasks. Our code, models, training data and evaluation set are available at https://github.com/X-PLUG/mPLUG-DocOwl.